Regression with limited dependent variables

- Professor Bernard Fingleton

Regression with limited dependent variables

- Whether a mortgage application is accepted or denied
- Decision to go on to higher education
- Whether or not foreign aid is given to a country
- Whether a job application is successful
- Whether or not a person is unemployed
- Whether a company expands or contracts

Regression with limited dependent variables

- In each case, the outcome is binary
- We can treat the variable as a success (Y = 1) or failure (Y = 0)
- We are interested in explaining the variation across people, countries or companies etc in the probability of success, p = prob(Y =1)
- Naturally we think of a regression model in which Y is the dependent variable

Regression with limited dependent variables

- But the dependent variable *Y* and hence the errors are not what is assumed in ‘normal’ regression
  - Continuous range
  - Constant variance (homoscedastic)
- With individual data, the *Y* values are 1(success) and 0(failure)
  - the observed data for N individuals are discrete values 0,1,1,0,1,0, etc……not a continuum
  - The variance is not constant (heteroscedastic)

Bernoulli distribution

probability of a success ($Y = 1$) is p

probability of failure ($Y = 0$) is $1 - p = q$

$E(Y) = p$

$\operatorname{var}(Y) = p(1 - p)$

as p_i varies for i= 1,...,N individuals

then both mean and variance vary

$E(Y_i) = p_i$

$\operatorname{var}(Y_i) = p_i(1 - p_i)$

regression explains variation in $E(Y_i) = p_i$ as a function of

some explanatory variables

$E(Y_i) = f(X_{1i}, ..., X_{Ki})$

but the variance is not constant as $E(Y_i)$ changes

whereas in OLS regression, we assume only the mean

varies as X varies, and the variance remains constant

The linear probability model

this is a linear regression model

$$Y_i = b_0 + b_1 X_{1i} + ...b_K X_{Ki} + e_i$$

$$\Pr(Y_i = 1 \big| X_{1i}, ..., X_{Ki}) = b_0 + b_1 X_{1i} + ...b_K X_{Ki}$$

b_1 is the change in the probability that $Y = 1$ associated with a unit change in X_1, holding constant $X_2 ....X_K$, etc

This can be estimated by OLS but

Note that since $\text{var}(Y_i)$ is not constant, we need to allow for heteroscedasticity in t, F tests and confidence intervals

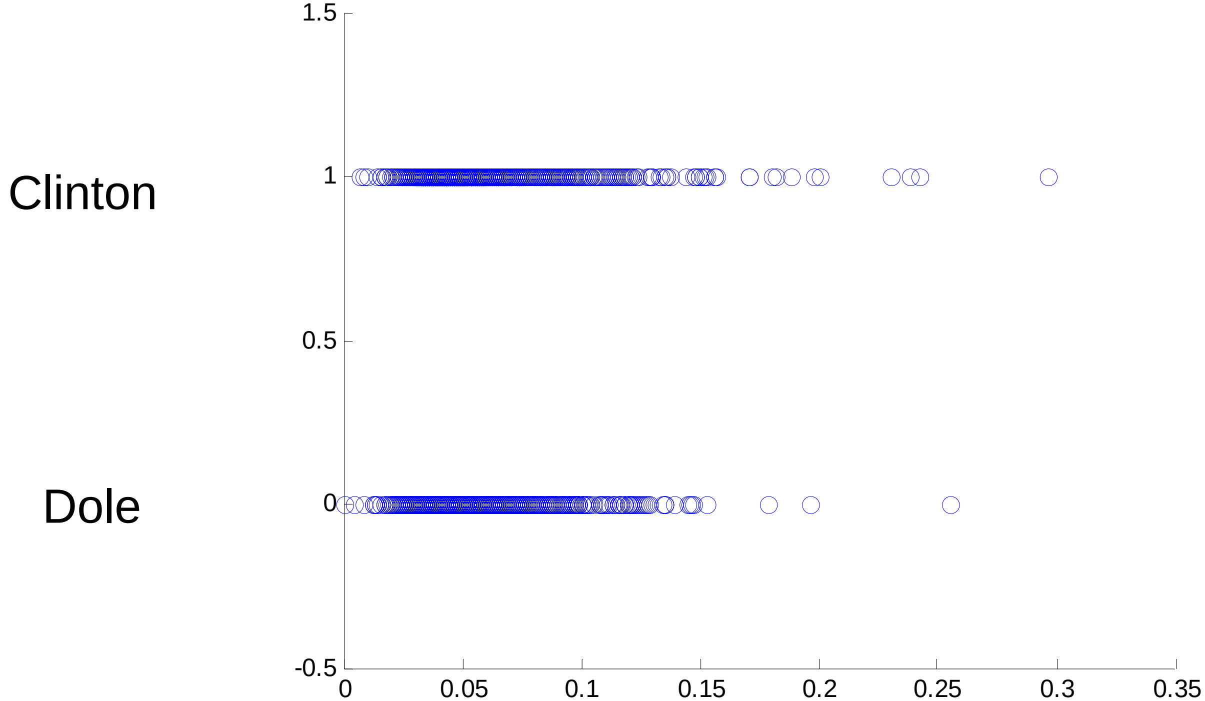
The linear probability model

- 1996 Presidential Election
- 3,110 US Counties
- binary *Y* with 0=Dole, 1=Clinton

The linear probability model

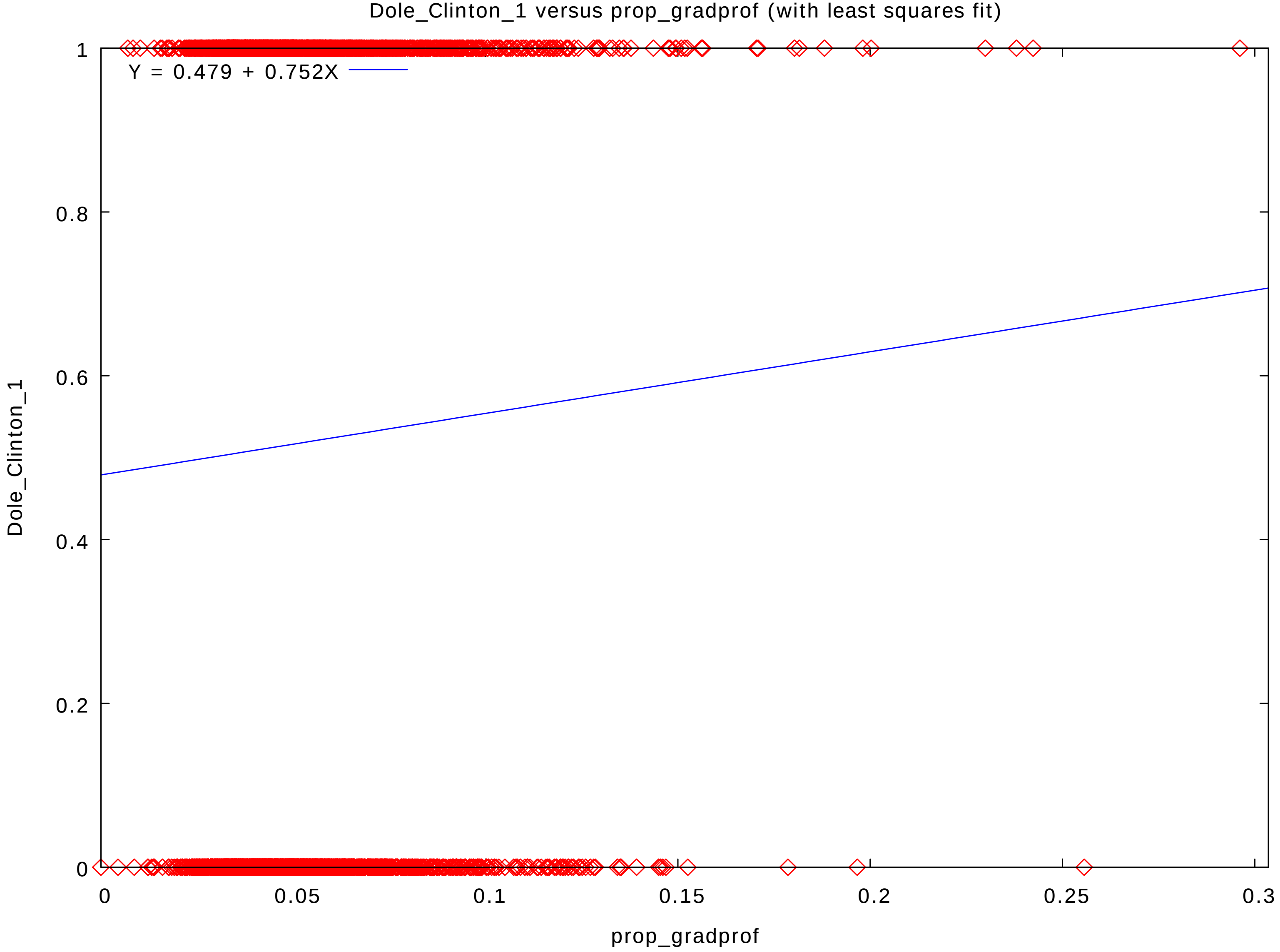
```
Ordinary Least-squares Estimates
R-squared      =    0.0013
Rbar-squared   =    0.0010
sigma^2        =    0.2494
Durbin-Watson  =    0.0034
Nobs, Nvars    =   3110,     2
***************************************************************
Variable         Coefficient      t-statistic    t-probability
Constant            0.478917        21.962788         0.000000
prop-gradprof       0.751897         2.046930         0.040749

prop-gradprof = pop with grad/professional degrees as a proportion of
educated (at least high school education)
```

The linear probability model



prop-gradprof = pop with grad/professional degrees as a proportion of educated (at least high school education)

The linear probability model



The linear probability model

- Limitations
- The predicted probability exceeds 1 as X becomes large

$$\hat{Y} = 0.479 + 0.752X$$

$$\text{if } X > 0.693 \text{ then } \hat{Y} > 1$$

$$X = 0 \text{ gives } \hat{Y} = 0.479$$

$$\text{if } X < 0 \text{ possible, then}$$

$$X < -0.637 \text{ gives } \hat{Y} < 0$$

Solving the problem

- We adopt a nonlinear specification that forces the dependent proportion to always lie within the range 0 to 1
- We use cumulative probability functions (cdfs) because they produce probabilities in the 0 1 range
- Probit
  - Uses the standard normal cdf
- Logit
  - Uses the logistic cdf

Probit regression

$\Phi(z) =$ area to left of z in standard normal distribution

$\Phi(-1.96) = 0.025$

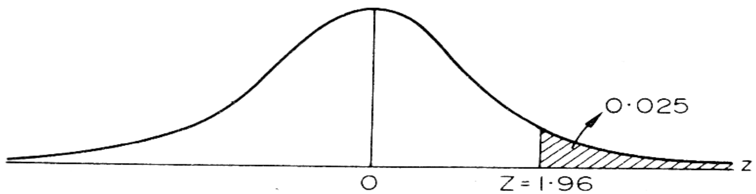
$\Phi(0) = 0.5$

$\Phi(1) = 0.84$

$\Phi(3.0) = 0.999$

we can put any value for z from -∞ to +∞, and the outcome is $0 < p = \Phi(z) < 1$

Table 1. Areas under the Normal Curve



Example

$$Z = \frac{X - \mu}{\sigma}$$

$$P(Z > 1{\cdot}96) = {\cdot}0250$$

z	·00	·01	·02	·03	·04	·05	·06	·07	·08	·09
0·0	·5000	·4960	·4920	·4880	·4840	·4801	·4761	·4721	·4681	·4641
0·1	·4602	·4562	·4522	·4483	·4443	·4404	·4364	·4325	·4286	·4247
0·2	·4207	·4168	·4129	·4090	·4052	·4013	·3974	·3936	·3897	·3859
0·3	·3821	·3783	·3745	·3707	·3669	·3632	·3594	·3557	·3520	·3483
0·4	·3446	·3409	·3372	·3336	·3300	·3264	·3228	·3192	·3156	·3121
0·5	·3085	·3050	·3015	·2981	·2946	·2912	·2877	·2843	·2810	·2776
0·6	·2743	·2709	·2676	·2643	·2611	·2578	·2546	·2514	·2483	·2451
0·7	·2420	·2389	·2358	·2327	·2296	·2266	·2236	·2206	·2177	·2148
0·8	·2119	·2090	·2061	·2033	·2005	·1977	·1949	·1922	·1894	·1867
0·9	·1841	·1814	·1788	·1762	·1736	·1711	·1685	·1660	·1635	·1611
1·0	·1587	·1562	·1539	·1515	·1492	·1469	·1446	·1423	·1401	·1379
1·1	·1357	·1335	·1314	·1292	·1271	·1251	·1230	·1210	·1190	·1170
1·2	·1151	·1131	·1112	·1093	·1075	·1056	·1038	·1020	·1003	·0985
1·3	·0968	·0951	·0934	·0918	·0901	·0885	·0869	·0853	·0838	·0823
1·4	·0808	·0793	·0778	·0764	·0749	·0735	·0721	·0708	·0694	·0681
1·5	·0668	·0655	·0643	·0630	·0618	·0606	·0594	·0582	·0571	·0559
1·6	·0548	·0537	·0526	·0516	·0505	·0495	·0485	·0475	·0465	·0455
1·7	·0446	·0436	·0427	·0418	·0409	·0401	·0392	·0384	·0375	·0367
1·8	·0359	·0351	·0344	·0336	·0329	·0322	·0314	·0307	·0301	·0294
1·9	·0287	·0281	·0274	·0268	·0262	·0256	·0250	·0244	·0239	·0233
2·0	·0228	·0222	·0217	·0212	·0207	·0202	·0197	·0192	·0188	·0183
2·1	·0179	·0174	·0170	·0166	·0162	·0158	·0154	·0150	·0146	·0143
2·2	·0139	·0136	·0132	·0129	·0125	·0122	·0119	·0116	·0113	·0110
2·3	·0107	·0104	·0102	·0099	·0096	·0094	·0091	·0089	·0087	·0084
2·4	·0082	·0080	·0078	·0075	·0073	·0071	·0069	·0068	·0066	·0064
2·5	·0062	·0060	·0059	·0057	·0055	·0054				

Probit regression

$$\Pr(Y=1 \mid X_1, X_2) = \Phi(b_0 + b_1 X_1 + b_2 X_2)$$

$e.g.$

$$b_0 = -1.6, b_1 = 2, b_2 = 0.5$$

$$X_1 = 0.4, X_2 = 1$$

$$z = b_0 + b_1 X_1 + b_2 X_2 = -1.6 + 2x0.4 + 0.5x1 = -0.3$$

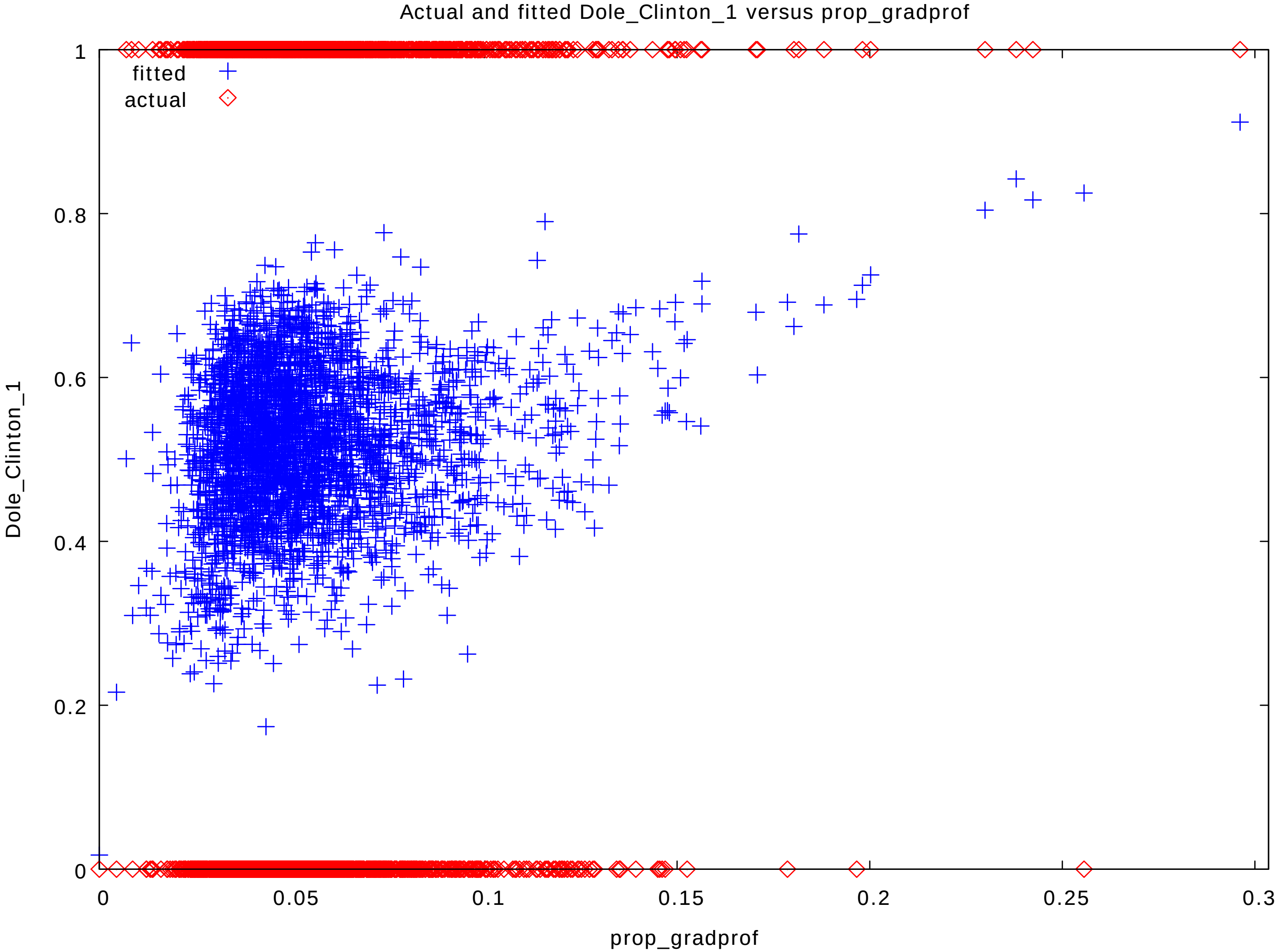
$$\Pr(Y=1 \mid X_1, X_2) = \Phi(-0.3) = 0.38$$

Probit regression

Model 9: Probit estimates using the 3110 observations 1-3110
Dependent variable: Dole_Clinton_1

VARIABLE	COEFFICIENT	STDERROR	T STAT	SLOPE (at mean)
const	-2.11372	0.215033	-9.830	
prop_gradprof	9.35232	1.32143	7.077	3.72660
log_urban	15.9631	5.64690	2.827	6.36078
prop_highs	3.07148	0.310815	9.882	1.22389

Probit regression



Probit regression

- Interpretation
- The slope of the line is not constant
- As the proportion of graduate professionals goes from 0.1 to 0.3, the probability of Y=1 (Clinton) goes from 0.5 to 0.9
- As the proportion of graduate professionals goes from 0.3 to 0.5 the probability of Y=1 (Clinton) goes from 0.9 to 0.99

Probit regression

- Estimation
- The method is maximum likelihood (ML)
- The likelihood is the joint probability given specific parameter values
- Maximum likelihood estimates are those parameter values that maximise the probability of drawing the data that are actually observed

Probit regression

$\Pr(Y_i = 1)$ conditional on $X_{1i}, \ldots, X_{Ki}$ is $p_i = \Phi(b_0 + b_1 X_{1i} + \ldots b_K X_{Ki})$

$\Pr(Y_i = 0)$ conditional on $X_{1i}, \ldots, X_{Ki}$ is $1 - p_i = 1 - \Phi(b_0 + b_1 X_{1i} + \ldots b_K X_{Ki})$

y_i is the value of Y_i observed for individual i

for i'th individual, $\Pr(Y_i = y_i)$ is $p_i^{y_i}(1 - p_i)^{1-y_i}$

for $i = 1, .., n$, joint likelihood is $L = \prod_i \Pr(Y_i = y_i) = \prod_i p_i^{y_i}(1 - p_i)^{1-y_i}$

$$L = \prod_i \Pr(Y_i = y_i) = \prod_i \left[\Phi(b_0 + b_1 X_{1i} + \ldots b_K X_{Ki})\right]^{y_i} \left[1 - \Phi(b_0 + b_1 X_{1i} + \ldots b_K X_{Ki})\right]^{1-y_i}$$

log likelihood is

$$\ln L = \sum_i \left[y_i \ln\left\{\Phi(b_0 + b_1 X_{1i} + \ldots b_K X_{Ki}\right\} + (1 - y_i)\ln\left\{1 - \Phi(b_0 + b_1 X_{1i} + \ldots b_K X_{Ki}\right\}\right]$$

we obtain the values of $b_0, b_1, \ldots, b_K$ that give the maximum value of $\ln L$

Hypothetical binary data

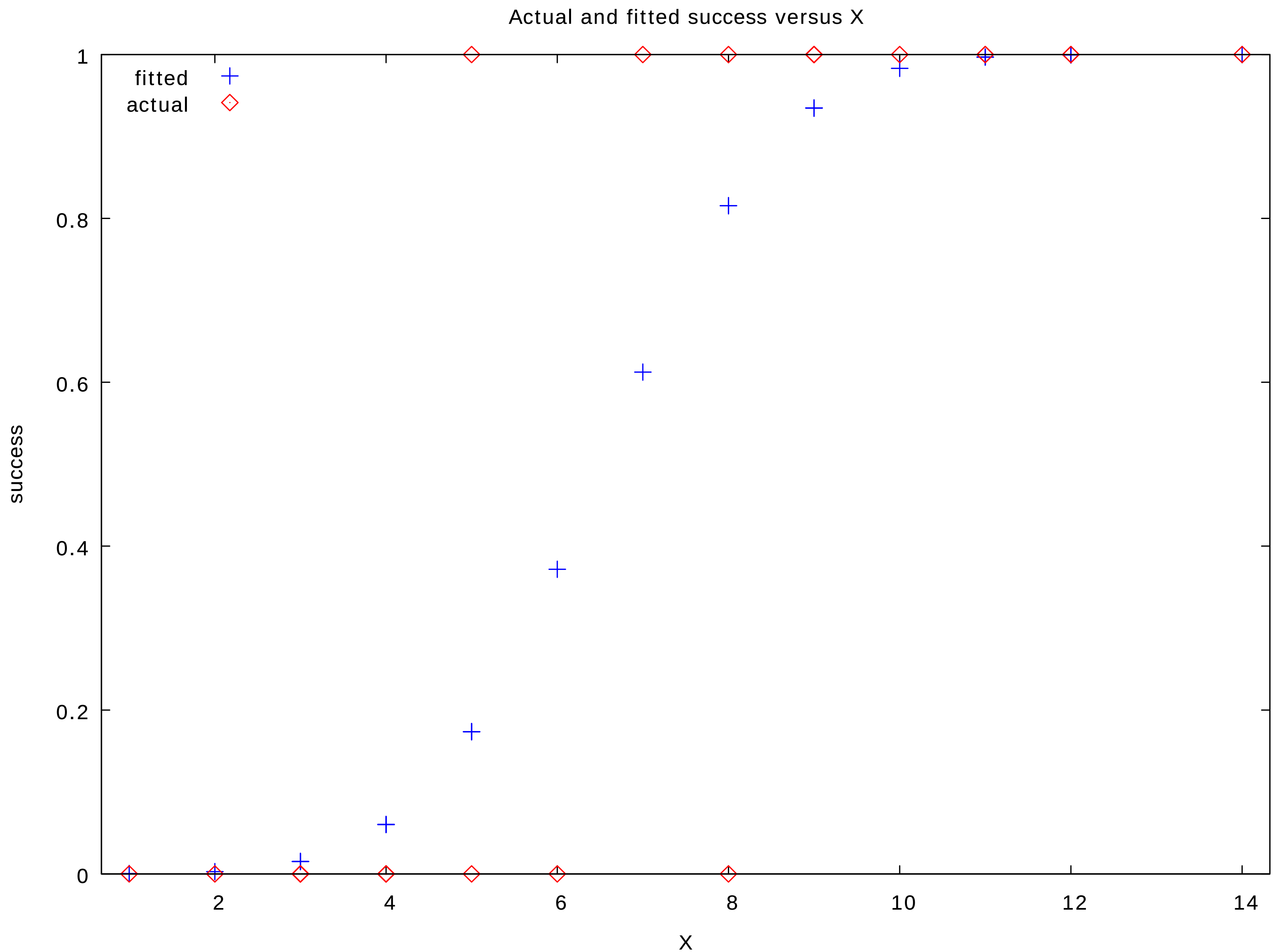
success	X
1	10
0	2
0	3
1	9
1	5
1	8
0	4
0	5
1	11
1	12
0	3
0	4
1	12
0	8
1	14
0	3
1	11
1	9
0	4
0	6
1	7
1	9
0	3
0	1

```
Iteration 0: log likelihood = -13.0598448595
Iteration 1: log likelihood = -6.50161713610
Iteration 2: log likelihood = -5.50794602456
Iteration 3: log likelihood = -5.29067548323
Iteration 4: log likelihood = -5.26889753239
Iteration 5: log likelihood = -5.26836878709
Iteration 6: log likelihood = -5.26836576121
Iteration 7: log likelihood = -5.26836575008

Convergence achieved after 8 iterations

Model 3: Probit estimates using the 24 observations 1-24
Dependent variable: success

      VARIABLE       COEFFICIENT        STDERROR      T STAT       SLOPE
                                                                  (at mean
  const                -4.00438          1.45771     -2.747
  X                     0.612845         0.218037     2.811      0.241462
```

Model 3: Probit estimates using the 24 observations 1-24
Dependent variable: success

VARIABLE	COEFFICIENT	STDERROR	T STAT	SLOPE (at mean)
const	-4.00438	1.45771	-2.747	
X	0.612845	0.218037	2.811	0.241462

If X = 6, probit = $\Phi(-4.00438 + 0.612845 * 6) = \Phi(-0.32731) = 0.45$

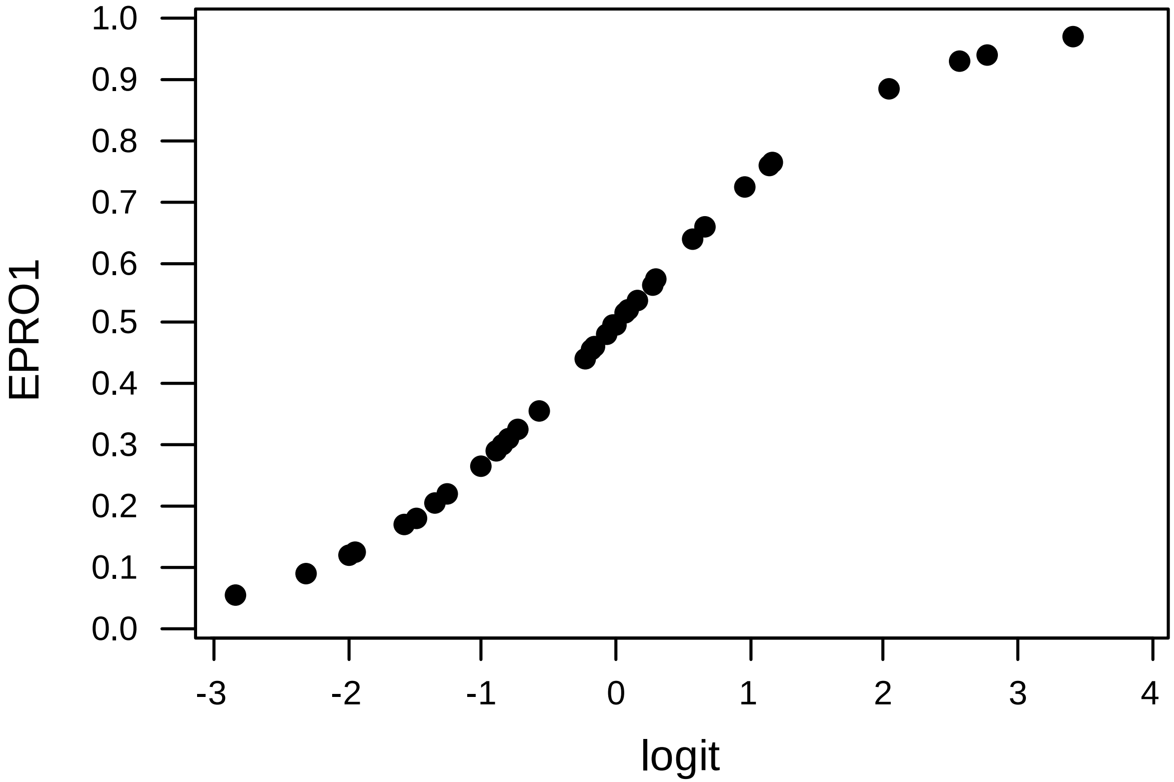


Actual and fitted success versus X
fitted
actual
success
X
1
0.8
0.6
0.4
0.2
0
2
4
6
8
10
12
14

Logit regression

- Based on logistic cdf
- This looks very much like the cdf for the normal distribution
- Similar results
- The use of the logit is often a matter of convenience, it was easier to calculate before the advent of fast computers

Logistic function



f(z)

z

$$\text{Prob success} = \frac{e^{Z}}{1+e^{Z}} = \left[1+e^{-z}\right]^{-1}$$

$$\text{Prob fail} = 1 - \frac{e^{Z}}{1+e^{Z}} = \frac{1+e^{Z}}{1+e^{Z}} - \frac{e^{Z}}{1+e^{Z}} = \frac{1}{1+e^{Z}} = 1-\left[1+e^{-z}\right]^{-1}$$

$$\text{odds ratio} = \frac{\text{Prob success}}{\text{Prob fail}} = \frac{e^{Z}}{1+e^{Z}} \frac{1+e^{Z}}{1} = e^{Z}$$

$$\text{log odds ratio} = z$$

$$z = b_0 + b_1 X_1 + b_2 X_2 + \ldots.. + b_k X_k$$

Logistic function

$$p_i = b_0 + b_1 X$$

p plotted against X is a straight line with p <0 and >1 possible

$$p_i = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

p_i plotted against X gives s-shaped logistic curve

so $p_i > 1$ and $p_i < 0$ impossible

equivalently

$$\ln\left\{\frac{p_i}{1 - p_i}\right\} = b_0 + b_1 X$$

this is the equation of a straight line, so

$\ln\left\{\frac{p_i}{1 - p_i}\right\}$ plotted against X is linear

Estimation - logit

X is fixed data, so we choose $b_0, b_1,$ hence p_i

$$p_i = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

so that the likelihood is maximized

Logit regression

$z = b_0 + b_1 X_{1i} + ... b_K X_{Ki}$

$\Pr(Y_i = 1)$ conditional on $X_{1i}, ..., X_{Ki}$ is $p_i = [1 + \exp(-z)]^{-1}$

$\Pr(Y_i = 0)$ conditional on $X_{1i}, ..., X_{Ki}$ is $1 - p_i = 1 - [1 + \exp(-z)]^{-1}$

y_i is the value of Y_i observed for individual i

for i'th individual, $\Pr(Y_i = y_i)$ is $p_i^{y_i} (1 - p_i)^{1 - y_i}$

for $i = 1, .., n,$ joint likelihood is $L = \prod_i \Pr(Y_i = y_i) = \prod_i p_i^{y_i} (1 - p_i)^{1 - y_i}$

$$L = \prod_i \Pr(Y_i = y_i) = \prod_i \left[[1 + \exp(-z)]^{-1} \right]^{y_i} \left[1 - [1 + \exp(-z)]^{-1} \right]^{1 - y_i}$$

log likelihood is

$$\ln L = \sum_i \left[y_i \ln \left\{ [1 + \exp(-z)]^{-1} \right\} + (1 - y_i) \ln \left\{ 1 - [1 + \exp(-(z)]^{-1} \right\} \right]$$

we obtain the values of $b_0, b_1, ..., b_K$ that give the maximum value of $\ln L$

Estimation

- maximum likelihood estimates of the parameters using an iterative algorithm

Estimation

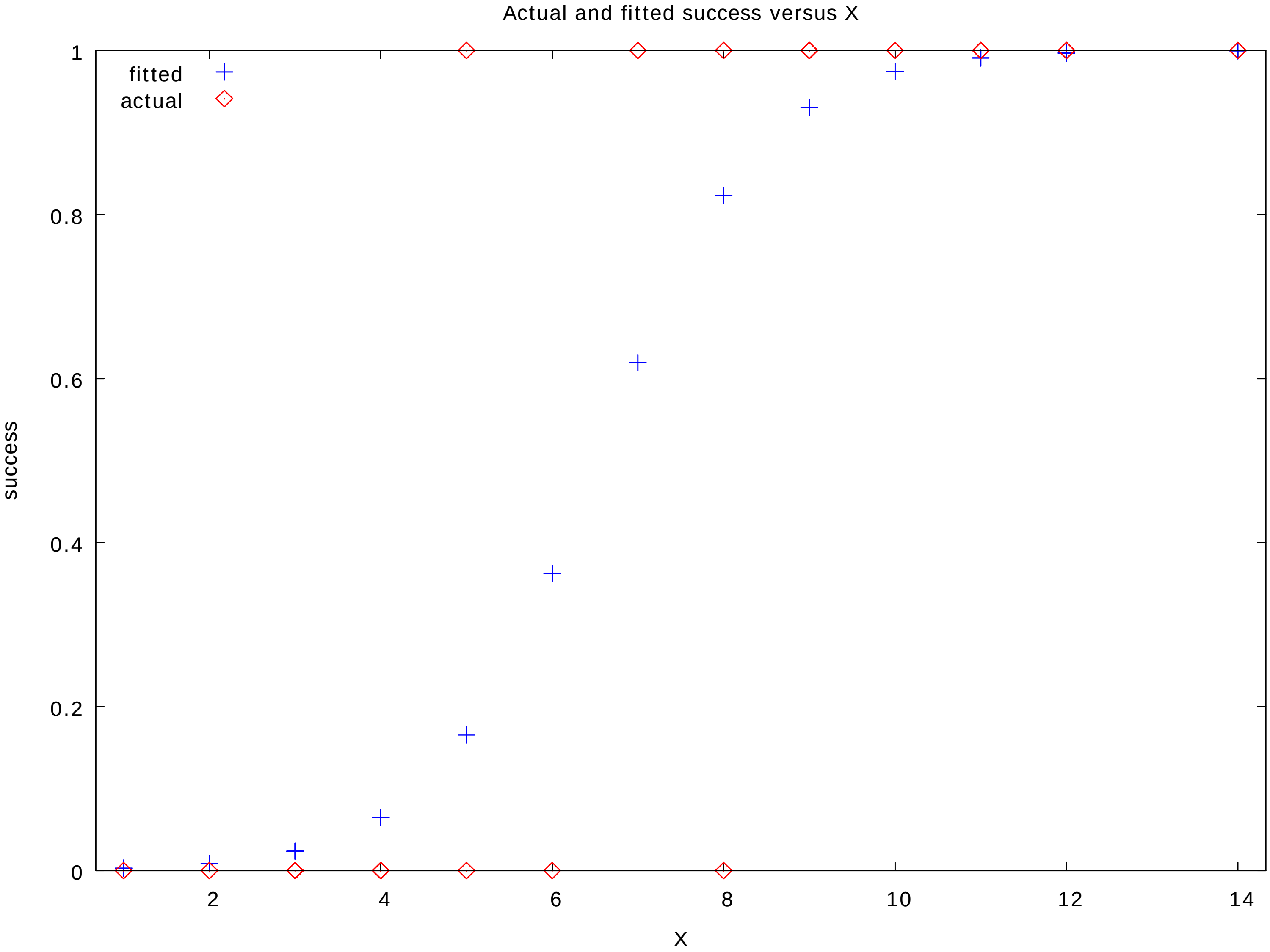
```
Iteration 0: log likelihood = -13.8269570846
Iteration 1: log likelihood = -6.97202524093
Iteration 2: log likelihood = -5.69432863365
Iteration 3: log likelihood = -5.43182376684
Iteration 4: log likelihood = -5.41189406278
Iteration 5: log likelihood = -5.41172246346
Iteration 6: log likelihood = -5.41172244817

Convergence achieved after 7 iterations

Model 1: Logit estimates using the 24 observations 1-24
Dependent variable: success
```

VARIABLE	COEFFICIENT	STDERROR	T STAT	SLOPE (at mean)
const	-6.87842	2.74193	-2.509	
X	1.05217	0.408089	2.578	0.258390

VARIABLE	COEFFICIENT	STDERROR	T STAT	SLOPE (at mean)
const	-6.87842	2.74193	-2.509	
X	1.05217	0.408089	2.578	0.258390

If X = 6, logit = -6.87842 + 1.05217*6 =-0.5654 = ln(p/(1-p)

P = exp(-0.5654)/{1+exp(-0.5654)} = 1/(1+exp(0.5654)) = 0.362299


Actual and fitted success versus X
fitted
actual
success
X
0
0.2
0.4
0.6
0.8
1
2
4
6
8
10
12
14

Modelling proportions and percentages

The linear probability model

consider the following individual data for Y and X

$Y = 0, 0, 1, 0, 0, 1, 0, 1, 1, 1$

$X = 1, 1, 2, 2, 3, 3, 4, 4, 5, 5$

constant $= 1, 1, 1, 1, 1, 1, 1, 1, 1, 1$

$\hat{Y} = -0.1 + 0.2X$ is the OLS estimate

Notice that the X values for individuals 1 and 2 are identical, likewise 3 and 4 and so on

If we group the identical data, we have a set of proportions

p = 0/2, 1/2, 1/2, 1/2, 1 = 0, 0.5, 0.5, 0.5, 1

X = 1,2,3,4,5

$\hat{p} = -0.1 + 0.2X$ is the OLS estimate

The linear probability model

- When n individuals are identical in terms of the variables explaining their success/failure
- Then we can group them together and explain the proportion of ‘successes’ in n trials
  - This data format is often important with say developing country data, where we know the proportion, or % of the population in each country with some attribute, such as the % of the population with no schooling
  - And we wish to explain the cross country variations in the %s by variables such as GDP per capita or investment in education, etc

Regression with limited dependent variables

- With individual data, the values are 1(success) and 0(failure) and p is the probability that $Y = 1$
  - the observed data for N individuals are discrete values 0,1,1,0,1,0, etc……not a continuum
- With grouped individuals the proportion p is equal to the number of successes Y in n trials (individuals)
- So the range of Y is from 0 to n
- The possible Y values are discrete, $0,1,2,\ldots,n$, and confined to the range 0 to n.
- The proportions p are confined to the range 0 to 1

Modelling proportions

Proportion (*Y/n*)	Continuous response
5/10 = 0.5	11.32
1/3 = 0.333	17.88
6/9 = 0.666	3.32
1/10 = 0.1	11.76
7/20 = 0.35	1.11
1/2 = 0.5	0.03

Binomial distribution

the moments of the number of successes Y_i

n_i trials, each independent, $i = 1, \ldots, N$

p_i is the probability of a success in each trial

$E(Y_i) = n_i p_i$

$\mathrm{var}(Y_i) = n_i p_i (1 - p_i)$

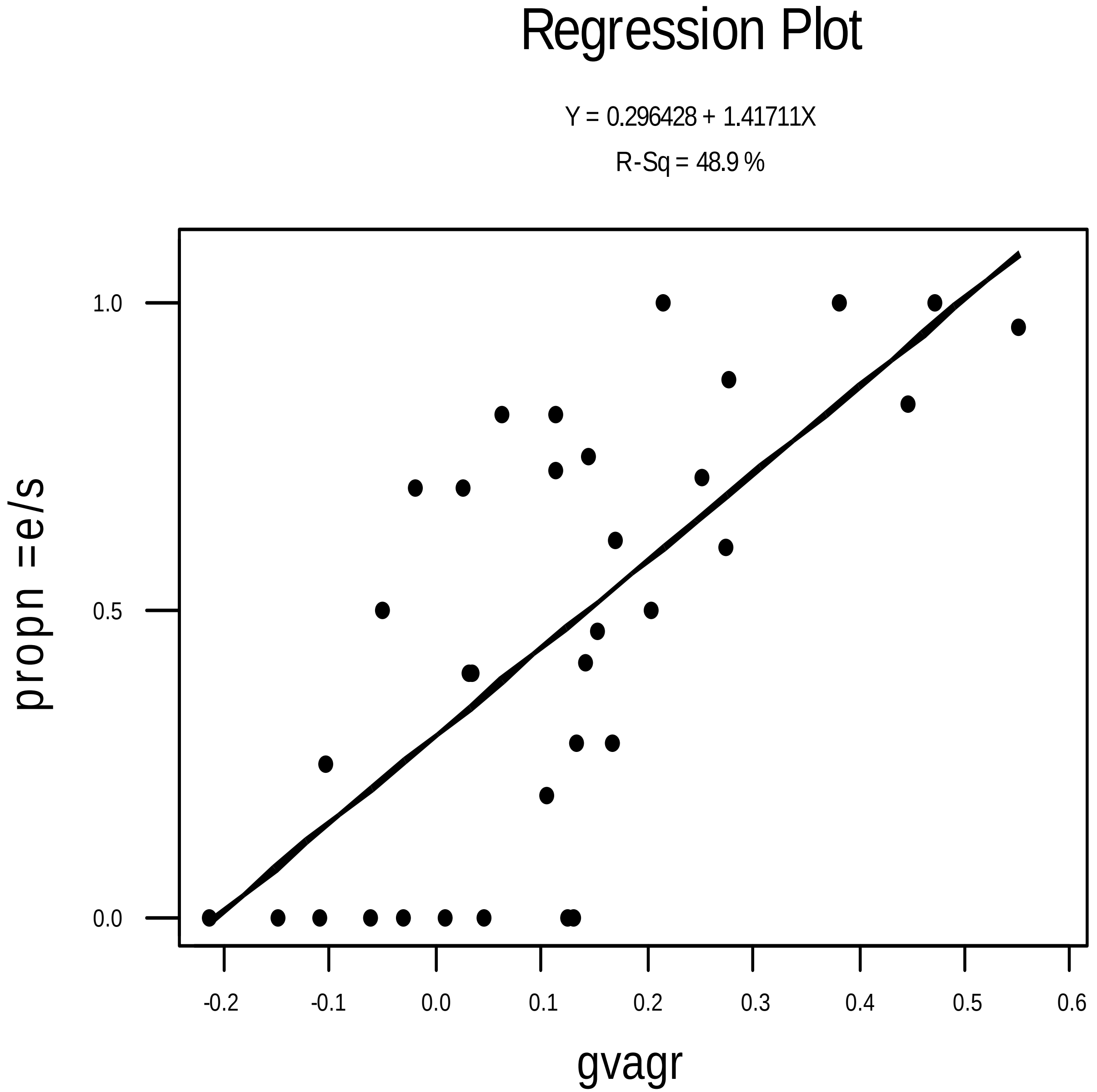
the variance is not constant, but depends on n_i and p_i

$Y_i \sim B(n_i, p_i)$

Data

Region	Output growth	survey of startup firms		
	q	starts(n)	expanded(Y)	propn $=Y/n$
Cleveland,Durham	0.169211	13	8	0.61538
Cumbria	0.471863	34	34	1.00000
Northhumberland	0.044343	10	0	0.00000
Humberside	0.274589	15	9	0.60000
N Yorks	0.277872	16	14	0.87500

The linear probability model



OLS regression with proportions

```
y = 0.296 + 1.42 x

Predictor         Coef      StDev        T       P
Constant       0.29643    0.05366     5.52   0.000
x               1.4171     0.2559     5.54   0.000

S = 0.2576     R-Sq = 48.9%     R-Sq(adj) = 47.3%
```

Fitted values = y = 0.296 + 1.42 x

	0.48310
Negative proportion	-0.00576
Proportion > 1	1.07892
	0.58634
	0.25346

Grouped Data

Region	Output growth	survey of startup firms		
	q	starts(n)	expanded(Y)	propn $=Y/n$
Cleveland,Durham	0.169211	13	8	0.61538
Cumbria	0.471863	34	34	1.00000
Northhumberland	0.044343	10	0	0.00000
Humberside	0.274589	15	9	0.60000
N Yorks	0.277872	16	14	0.87500

Proportions and counts

$$\ln(p_i/(1-p_i)) = b_0 + b_1 X$$

$$\ln(\hat{p}_i/(1-\hat{p}_i)) = \hat{b}_0 + \hat{b}_1 X$$

$$E(Y_i) = n_i p_i$$

$$\hat{Y}_i = n_i \hat{p}_i$$

n_i = size of sample i

$\hat{Y}_i$ = estimated expected number of ‘successes’ in sample i

Binomial distribution

For region i

$$\Pr ob(Y = y) = \frac{n!}{y!(n-y)!} p^{y} (1-p)^{n-y}$$

n = number of individuals

p = probability of a 'success'

Y = number of 'successes' in n individuals

Binomial distribution

For region i

$$\Pr ob(Y = y) = \frac{n!}{y!(n-y)!} p^{y}(1-p)^{n-y}$$

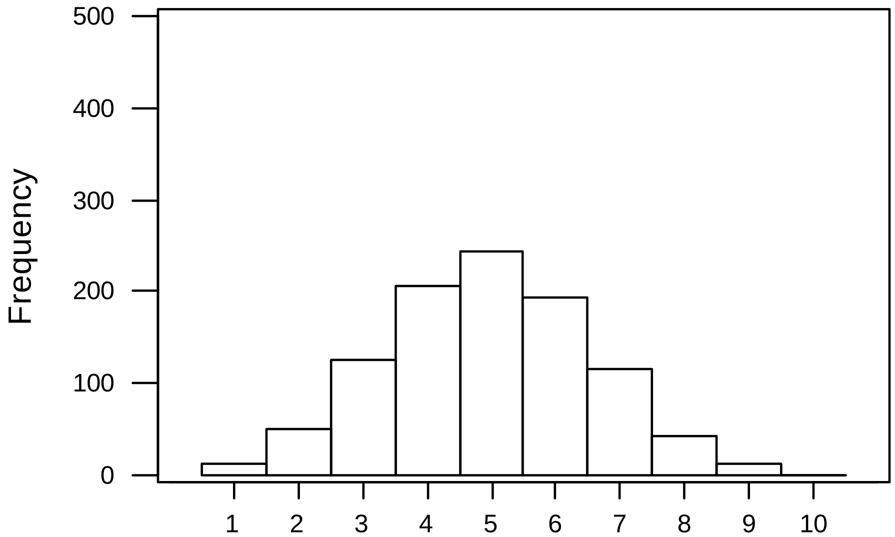
Example P =0.5, n = 10

$$\Pr ob(Y = 5) = \frac{10!}{5!(10-5)!} 0.5^{5}(0.5)^{5} = 0.2461$$

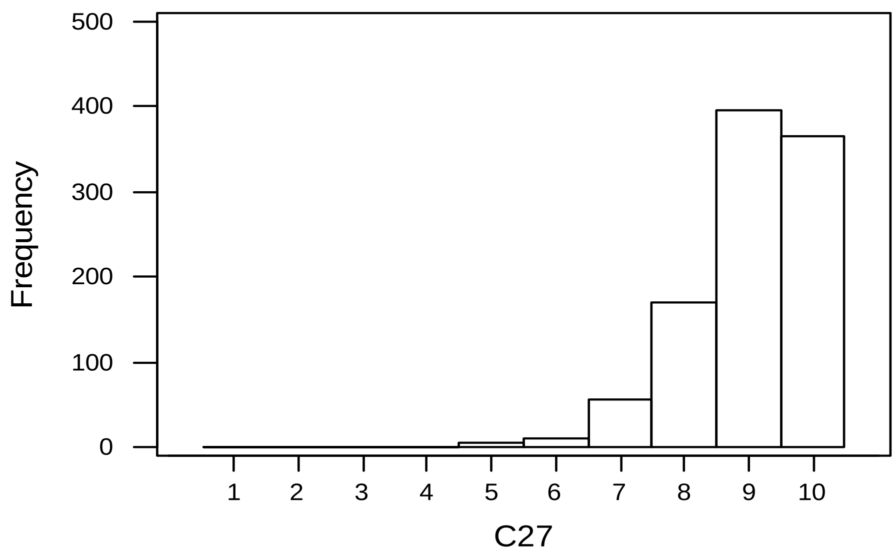
$$E(Y) = np = 5$$

$$\mathrm{var}(Y) = np(1-p) = 2.5$$

Y is B(10,0.5) E(Y)=np = 5 var(Y)=np(1-p)=2.5



Y is B(10,0.9) E(Y)=np = 9 var(Y)=np(1-p)=0.9



Maximum Likelihood – proportions

Assume the data observed are

$Y_1 = y_1 = 5$ successes from 10 trials and $Y_2 = y_2 = 9$ successes from 10 trials

what is the likelihood of these data given $p_1 = 0.5, p_2 = 0.9$?

$$\Pr ob(Y_1 = 5) = \frac{n_1!}{y_1!(n_1 - y_1)!}\; p_1^{y_1}(1 - p_1)^{n_1 - y_1}$$

$$\Pr ob(Y_1 = 5) = \frac{10!}{5!(10-5)!} 0.5^5 (0.5)^5 = 0.2461$$

$$\Pr ob(Y_2 = 9) = \frac{n_2!}{y_2!(n_2 - y_2)!}\; p_2^{y_2}(1 - p_2)^{n_2 - y_2}$$

$$\Pr ob(Y_2 = 9) = \frac{10!}{9!(10-9)!} 0.9^9 (0.1)^1 = 0.3874$$

likelihood of observing $y_1 = 5, y_2 = 9$ given $p_1 = 0.5, p_2 = 0.9$

$= 0.2461x0.3874 = 0.095$

However likelihood of observing $y_1 = 5, y_2 = 9$ given $p_1 = 0.1, p_2 = 0.8$

$= 0.0015x0.2684 = 0.0004$

Inference

Likelihood ratio/deviance

$$Y^2 = 2\ln(L_u / L_r) \sim \chi^2$$

L_u = likelihood of unrestricted model with k1 df

L_r = likelihood of restricted model with k2 df

k2 > k1

Restrictions placed on k2-k1 parameters
typically they are set to zero

Deviance

$$Ho: b_i = 0, i = 1, \ldots, (k2 - k1)$$

$$Y^2 = 2\ln(L_u / L_r) \sim \chi^2_{k2-k1}$$

$$E(Y^2) = k2 - k1$$

```
Iteration 7: log likelihood = -5.26836575008
```
= Lu

```
Convergence achieved after 8 iterations

Model 3: Probit estimates using the 24 observations 1-24
Dependent variable: success

     VARIABLE       COEFFICIENT        STDERROR      T STAT        SLOPE
                                                                 (at mean)
  const                -4.00438          1.45771      -2.747
  X                     0.612845         0.218037      2.811      0.241462


Model 4: Probit estimates using the 24 observations 1-24
Dependent variable: success

     VARIABLE       COEFFICIENT        STDERROR      T STAT        SLOPE
                                                                 (at mean)
  const                 0.000000         0.255832     -0.000


  Log-likelihood = -16.6355
```
=Lr

```
  Comparison of Model 3 and Model 4:

  Null hypothesis: the regression parameters are zero for the variables

    X
```

$$2\{Lu - Lr]= 2[-5.268 + 16.636] = 22.73$$

```
  Test statistic: Chi-square(1) = 22.7343, with p-value = 1.86014e-006

  Of the 3 model selection statistics, 0 have improved.
```

Iteration 3: log likelihood = -2099.98151495

Convergence achieved after 4 iterations

Model 2: Logit estimates using the 3110 observations 1-3110
Dependent variable: Dole_Clinton_1

VARIABLE	COEFFICIENT	STDERROR	T STAT	SLOPE (at mean)
const	-3.41038	0.351470	-9.703	
log_urban	25.4951	9.10570	2.800	6.36359
prop_highs	4.96073	0.508346	9.759	1.23820
prop_gradprof	15.1026	2.16268	6.983	3.76961

Model 3: Logit estimates using the 3110 observations 1-3110
Dependent variable: Dole_Clinton_1

VARIABLE	COEFFICIENT	STDERROR	T STAT	SLOPE (at mean)
const	-0.972329	0.184314	-5.275	
prop_highs	2.09692	0.360723	5.813	0.523414

Log-likelihood = -2136.12

Comparison of Model 2 and Model 3:

Null hypothesis: the regression parameters are zero for the variables

log_urban
prop_gradprof

Test statistic: Chi-square(2) = 72.2793, with p-value = 2.01722e-016

DATA LAYOUT FOR LOGISTIC REGRESSION

REGION	URBAN	SE/NOT SE	OUTPUT GROWTH	Y	n
Hants, IoW	suburban	SE	0.062916	9	11
Kent	suburban	SE	0.035541	4	10
Avon	suburban	not SE	0.133422	4	14
Cornwall, Devon	rural	not SE	0.141939	5	12
Dorset, Somerset	rural	not SE	0.145993	12	16
S Yorks	urban	not SE	-0.150591	0	11
W Yorks	urban	not SE	0.152066	7	15

Logistic Regression Table

Predictor	Coef	StDev	Z	P
Constant	-1.2132	0.3629	-3.34	0.001
gvagr	9.716	1.251	7.77	0.000
URBAN/SUBURBAN/RURAL				
suburban	-0.8464	0.2957	-2.86	0.004
urban	-1.3013	0.4760	-2.73	0.006
SOUTH-EAST/NOT SOUTH-EAST				
South-East	2.4411	0.3534	6.91	0.000

Log-Likelihood = -210.068

Testing variables

	Log-like	degrees of freedom
Prob = f(q)	-247.1	32
Prob = f(q,urban,SE)	-210.068	29

2*Difference = 74.064

74.064 > 7.81, the critical value equal to the upper 5% point of the chi-squared distribution with 3 degree of freedom
thus introducing URBAN/SUBURBAN/RURAL and SE/not SE causes a significant improvement in fit

Interpretation

- When the transformation gives a linear equation linking the dependent variable and the independent variables then we can interpret it in the normal way
- The regression coefficient is the change in the dependent variable per unit change in the independent variable, controlling for the effect of the other variables
- For a dummy variable or factor with levels, the regression coefficient is the change in the dependent variable associated with a shift from the baseline level of the factor

Interpretation

$$\ln(\hat{P}_i / (1 - \hat{P}_i))$$

Changes by 9.716 for a unit change in gvagr

by 2.441 as we move from not SE to SE counties

by -1.3013 as we move from RURAL to URBAN

by -0.8464 as we move from RURAL to SUBURBAN

Interpretation

The odds of an event = ratio of Prob(event) to Prob(not event)

The odds ratio is the ratio of two odds.

The logit link function means that parameter estimates are the exponential of the odds ratio (equal to the logit differences).

Interpretation

For example, a coefficient of zero would indicate that moving from a non SE to a SE location produces no change in the logit

Since exp(0) = 1, this would mean the (estimated) odds = Prob(expand)/Prob(not expand) do not change ie the odds ratio =1

In reality, since exp(2.441) = 11.49 the odds ratio is 11.49
The odds of SE firms expanding are 11.49 times the odds of non SE firms expanding

Interpretation

	param. est.	s.e.	t ratio	p-value	odds ratio	lower c.i.	upper c.i.
Constant	-1.2132	0.3629	-3.34	0.001			
gvagr	9.716	1.251	7.77	0.000	16584.51	1428.30	1.93E+05
RURAL/SUBURBAN/URBAN							
suburban	-0.8464	0.2957	-2.86	0.004	0.43	0.24	0.77
urban	-1.3013	0.4760	-2.73	0.006	0.27	0.11	0.69
SE/not SE							
SE	2.4411	0.3534	6.91	0.000	11.49	5.75	22.96

Note that the odds ratio has a 95% confidence interval
since 2.4411+1.96*0.3534 =3.1338
and 2.4411-1.96*0.3534 = 1.7484
and exp(3.1338)=22.96 , exp(1.7484) = 5.75
The 95% c.i.for the odds ratio is 5.75 to 22.96